引用格式: 张青露, 陈壹华. 基于特征共现的无监督哈希检索算法[J]. 微电子学与计算机, 2024, 41(5): 22-30.
ZHANG Q L, CHEN Y H. An unsupervised hash retrieval algorithm based on feature co-occurrence[J]. Microelectronics & Computer, 2024, 41(5): 22-30.
DOI: 10.19304/J.ISSN1000-7180.2023.0289

基于特征共现的无监督哈希检索算法

张青露, 陈壹华
(华南师范大学 软件学院, 广东 佛山 528225)

摘 要: 现有无监督哈希检索算法的关注点在于哈希映射过程中的信息损失以及生成哈希的质量问题, 忽略了图像特征本身对检索精度的影响。为进一步提高检索的精度, 提出一种改进的基于特征共现的无监督哈希检索算法(Unsupervised Hash retrieval algorithm based on Feature Co-occurrence, UHFC)。该算法共分为两个阶段: 深度特征提取和无监督哈希生成。为提高图像特征的质量, UHFC 在卷积神经网络 (Convolutional Neural Network, CNN) 结构的最后一层卷积后引入了共现层, 用来提取特征之间的依赖关系。并用共现激活值的均值来表示共现程度, 解决原共现操作存在相同两个通道的共现值不一致的问题; 接着, 在特征融合部分 UHFC 设计一种适用于共现特征融合的, 结合空间注意力机制的注意特征融合方法(Attention Feature Fusion method based on Spatial attention, AFF-S)。通过注意力机制自主学习共现特征与深度特征融合的权重, 降低特征融合过程中背景因素的干扰, 提高最终图像特征的表达能力。最后, 根据最优传输策略, UHFC 采用双半分布哈希编码对图像特征到哈希码的映射过程进行监督, 并在哈希层后添加一层分类层通过 KL 损失进一步提高哈希码所包含的图片信息, 整个训练过程中无需数据集的标注, 实现无监督哈希的生成。实验表明, UHFC 对哈希编码质量改善较好, 在 Flickr25k 和 Nus-wide 数据集上其平均均值精度(mean Average Precision, mAP)分别达到了 87.8% 和 82.8%, 相比于 baseline 方法分别提高了 2.1% 与 1.2%, 效果明显。

关键词: 图像检索; 注意特征融合; 共现特征; 无监督哈希
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 1000-7180(2024)05-0022-09

An unsupervised hash retrieval algorithm based on feature co-occurrence

ZHANG Qinglu, CHEN Yihua
(School of Software, South China Normal University, Foshan 528225, China)

Abstract: The existing algorithms of unsupervised hash retrieval focus on the information loss in the process of hash mapping and the quality of hash generation, but ignore the impact of image features on the retrieval accuracy. In order to further improve the retrieval accuracy, this paper proposes an improved Unsupervised Hash retrieval algorithm based on Feature Co-occurrence (UHFC), which is divided into two stages: deep feature extraction and unsupervised hash generation. In order to improve the quality of image features, UHFC introduces a co-occurrence layer after the last convolution layer of Convolutional Neural Network (CNN) structure to extract the dependency relationship between features. The mean value of co-occurrence activation value is used to represent the degree of co-occurrence to solve the problem of inconsistent co-present value of the same two channels in the original co-occurrence operation. Then, in the feature fusion part of UHFC, an Attentional Feature Fusion method based on Spatial attention(AFF-S) mechanism is designed for co-occurrence feature fusion. By self-learning the weight of co-occurrence feature and depth feature fusion by attention mechanism, the interference of background factors in the process of feature fusion is reduced, and the expressive ability of final image

收稿日期: 2023-04-10; **修回日期:** 2023-05-12
基金项目: 广东省基础与应用基础研究基金(2021A1515011171)

features is improved. Finally, according to the optimal transmission strategy, UHFC adopts Bi-half distributed hash coding to supervise the mapping process of image features to hash code, and adds a classification layer after the hash layer to further improve the image information contained in the hash code through KL loss. In the whole training process, no data set labeling is required to realize the generation of unsupervised hash. Experiments have shown that UHFC better improve quality of hash code, in Flickr25k and Nus - wide data sets its mean Average Precision (mAP) reached 87.8% and 82.8% respectively, compared to the baseline method is increased by 2.1% and 1.2%, respectively, effect is obvious.

Key words：image retrieval；attention feature fusion；co-occurrence feature；unsupervised hashing

1　引言

随着互联网技术的不断发展，各种信息和数据层出不穷，图像已经成为互联网文本之外信息的主要载体之一。面对网络上数以亿计的图像数据，如何快速准确地检索到相似图像已成为一个非常重要和热门的研究课题之一。哈希图像检索算法提取图片的特征信息，并将该特征转换成一维二进制编码存储在数据库中，用户检索时通过比对检索图片哈希码与数据库哈希码之间的汉明距离来判断图片之间的相似性，获得检索结果。哈希检索由于其紧凑的二值表示和高效的汉明距离计算，成为大规模图像检索的主流。

Datar 等[1] 提出了局部敏感哈希(Locality-Sensitive Hashing, LSH)，将图像特征通过哈希函数映射成二进制序列，实现了在大型数据库的快速检索。但在 CNN 训练过程中，往往需要大量标注好的数据[2-4]，耗费大量人力物力，人们开始尝试使用无监督的方式对模型进行训练。Ma 等[5] 利用锚点数据与训练数据构成的相似性矩阵来作为监督信息对哈希的生成进行训练。Li 等[6] 设计了一个无参数的双半编码层，使用 Wasserstein 距离，使连续特征与最优离散分布对齐，最大限度地提高哈希信道信息容量。考虑到哈希映射的过程往往会丢失大量的特征信息，Gong 等[7] 采用位损失和量化误差来降低转换过程中的语义损失，并结合分类损失进一步提高哈希码所包含的图片信息。Luo 等[8] 利用类间相似度损失来提高哈希码的位独立性和鲁棒性。然而这些方法都忽略了图像特征本身对哈希码的影响。

传统的图像特征如颜色直方图、灰度共生矩阵、局部尺度不变特征(Scale Invariant Feature Transform, SIFT)等，由于特征计算算法固定，具有良好的稳定性，但在 ImageNet ILSVRC 比赛中，最低的错误率也在 26% 以上。卷积神经网络(Convolutional Neural Network, CNN)的出现，使得图像检索向更精确更快的方向发展。2012 年 AlexNet[9] 在 ImageNet ILSVRC 比赛中降低了 10% 的错误率，2015 年 ResNet[10] 将错误率降低至 9%。为了获得更加高效的图像描述符，一些研究人员尝试将不同层次或不同尺度的图像特征进行融合[11-12]，另一些学者尝试将图像传统特征与深度特征进行融合，以获得更加丰富的图像信息[13]。但是这些融合方式都没有充分利用到卷积层之间的相关信息。

图像特征共现最早由 Yang 等[14] 提出，用来表示图像中视觉特征的空间依赖性。Zhu 等[15] 将特征共现的计算方法应用在卷积神经网络中，对神经元检测到的视觉部分之间的共现进行编码，并在实验中证实了在卷积层的后几层增加共现操作可以有效提高最终深度特征的质量。Forcen 等[16] 在此基础上，提出了共现过滤器，通过计算各通道激活图在给定区域下特征间的共现值，使得最终的共现张量包含更多的空间信息。但这种方法计算出的共现值具有不对称性，即相同的两个通道间的共现值不同，使得生成的共现张量不能很好地表达特征之间的相互依赖程度。另一方面，文献 [15-16] 中均采用张量拼接的方式对提取到的共现特征以及深度特征进行融合，这增加了网络中的参数量，不利于模型的优化。尽管文献 [7] 中采用了 1×1 卷积来减少特征映射的数量，但会损失最终图像检索的精度。

特征融合是计算机视觉领域备受关注的领域，ResNet 在卷积层之间采用短跳跃连接的方式构造多个堆叠的残差块[2]，实现卷积前后特征图的融合。Inception 通过构造不同的卷积块实现不同尺度特征的融合[17]。同样，FPN 通过构造特征金字塔对多尺度的图像特征进行融合[18]。这些方法都只在融合的对象做改进，特征融合的方法仍然是张量的连接或相加，使得特征融合的效果受限。为解决这个问题，Dai 等[19] 提出了注意特征融合模块，通过引入多尺度注意力机制实现特征融合权重的自主学习，而由于共现特征本身对通道之间的依赖关系的表示，并不适用于采用多尺度注意力机制来进行融合权重的学习。

为了提高哈希图像检索框架中深度特征的质量，进而提高哈希检索的精度。受 Forcen 等[16] 的启发，本文在 CNN 特征提取网络的最后一层卷积层后增

加一层共现层，提取特征之间的依赖关系。针对 Forcen 等[16]计算的共现张量存在非对称性问题，本文采用激活值的均值来表示通道激活值之间的共现程度，在保证共现矩阵对称的情况下更好地表示特征之间的依赖程度。提取到的共现特征通过基于空间注意力机制的注意特征融合模块与原深度特征进行融合，降低融合过程中背景特征的干扰，丰富了哈希映射前特征所携带的信息。哈希映射过程中，为降低映射过程中的特征损失，本文采用 Li 等[6]提出的双半分布的哈希编码作为监督信息，使得生成的哈希码所携带的信息熵达到最大，并在哈希层后添加一层分类层通过 Kullback-Leibler(KL)损失函数，提高哈希码携带的图像信息。训练过程中，无需数据集的标注。

2 UHFC 算法

图 1 为 UHFC 的整体框架图，UHFC 共分为深度特征的提取和无监督哈希的生成两个阶段，其中深度特征提取包括共现特征提取和共现特征融合两个部分。在共现特征提取部分，训练图片经过 CNN 网络，获得最后一层卷积特征 A。为了获得不同通道下特征激活图之间的依赖关系，采用共现过滤器 F 提取到张量 A 不同通道特征的共现特征 C_T。在特征融合的部分，共现特征 C_T 与原深度特征 A 通过基于空间注意力机制的注意特征融合方法(AFF-S)获得最终的图像特征 A_C。然后，图像特征 A_C 经过池化层和全连接层生成哈希映射前的图像特征描述符 V。在无监督哈希生成阶段，UHFC 通过使得每一个哈希位上 1 和−1 的相近，构造了双半分布的哈希编码层 B，并通过双半分布哈希 B 与连续特征 V 之间均方损失 L_{mes} 对经过 Sign()函数的哈希 H 进行优化。最后为进一步提高哈希码质量，UHFC 在 V 与 H 后分别添加了一层分类层获得 K_1 和 K_2，通过 K_1 和 K_2 之间的 KL 损失 L_{KL} 对 V 进一步优化，实现无监督哈希的生成。

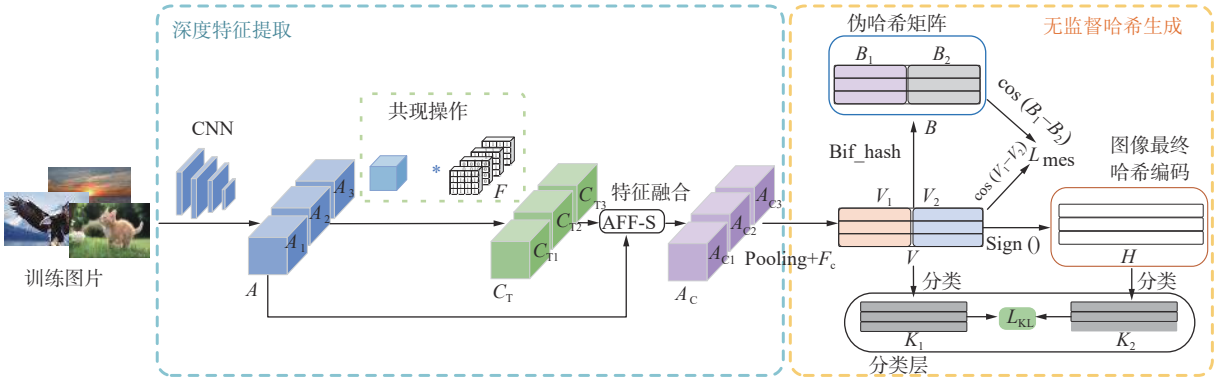

图 1 UHFC 整体框架

Fig. 1 UHFC Overall frame diagram

2.1 深度特征提取

2.1.1 共现特征提取

特征共现的概念来源于图像处理中的灰度共现矩阵，用来描述具有某种空间位置关系两个像素灰度的联合分布。Zhu 等[15]将这一概念引入到神经网络中，Forcen 等[16]在计算过程中加入了空间相关性，将特征共现定义为不同通道同一区域的两个激活值均大于给定激活值。

图 2 为图片经过卷积神经网络最后一层卷积的激活张量$A\in R^{M\times N\times D}$，$D$ 表示张量 A 的通道数，$M\times N$ 表示单个通道中激活图的尺寸。图 2 中，$a_{i,j}^{k}$表示第 k 通道第 i 行第 j 列位置的激活值，$a_{m,n}^{q}$同理。在给定区域 r 和阈值 s 下，激活值$a_{i,j}^{k}$和$a_{m,n}^{q}$的共现值 ι定义为：$a_{i,j}^{k}$和$a_{m,n}^{q}$同处于 $2r-1$ 的区域下且二者的激活值均大于 s，如式(1)所示。

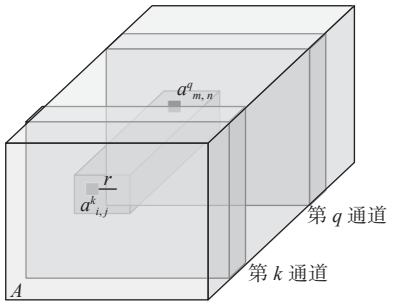

图 2 共现特征示意图

Fig. 2 Schematic diagram of co-occurrence feature

$$\iota(a_{i,j}^{k},a_{m,n}^{q})=\begin{cases}1 & |i-m|\leqslant r\cap|j-n|\leqslant r\cap a_{i,j}^{k}>s\cap a_{m,n}^{q}>s\\ 0 & \text{其他}\end{cases} \tag{1}$$

Forcen 等[16] 将第 k 通道，第 i 行第 j 列位置的共现张量值$C_{\mathrm{T}}(i,j,k)\in R^{M\times N\times D}$表示为$\iota(a_{i,j}^{k},a_{m,n}^{q})$与不同通道激活值$a_{m,n}^{q}$乘积的总和除以通道数，如式(2)所示：

$$C_{\mathrm{T}}(i,j,k)=\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{1}{D-1}\sum_{q=1}^{D}\iota(a_{i,j}^{k},a_{m,n}^{q})\cdot a_{m,n}^{q} \tag{2}$$

由于同一通道中的激活值不发生共现，因此，公式中的通道数为 $D-1$。但这种方法计算得到的共现张量不具有对称性，即当$\iota(a_{i,j}^{k},a_{m,n}^{q})$为 1 时，$C_{\mathrm{T}}(i,j,k)$在公式(2)中的$\iota(a_{i,j}^{k},a_{m,n}^{q})\cdot a_{m,n}^{q}$的值为$a_{m,n}^{q}$，而$C_{\mathrm{T}}(m,n,q)$在式(2)中$\iota(a_{i,j}^{k},a_{m,n}^{q})\cdot a_{m,n}^{q}$的值为$a_{i,j}^{k}$，即相同范围下的两个相同的通道的共现值不一致，使得生成的共现特征不能很好地表示不同通道下特征激活图之间的依赖关系。因此，本文对式(2)进行了改进，将第 k 通道，第 i 行第 j 列位置的共现张量值$C_{\mathrm{T}}(i,j,k)\in R^{M\times N\times D}$表示为$\iota(a_{i,j}^{k},a_{m,n}^{q})$乘以$a_{m,n}^{q}$和$a_{i,j}^{k}$均值的总和除以通道数，如式(3)所示：

$$C_{\mathrm{T}}(i,j,k)=\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{1}{D-1}\sum_{q=1}^{D}\iota(a_{i,j}^{k},a_{m,n}^{q})\cdot(a_{m,n}^{q}+a_{i,j}^{k})/2 \tag{3}$$

接着，参照 Forcen 等[16] 的做法，本文将式(1)中的 s 设置为张量 A 的均值，将共现发生的区域定义为尺寸为 $2r-1$ 的卷积核$F\in R^{D\times D\times W\times W}$，$D$ 为 A 的通道数，W 为卷积核尺寸。如式(4)所示，当通道数相同时，F 初始化为 0，其他情况下均设置为 1。式(4)中(a,b)代表 A 中两个通道数，(c,d)表示对应通道中激活值的位置。

$$F(a,b,c,d)\in R^{D\times D\times W\times W}=\begin{cases}0 & a=b\\ 1 & \text{其他}\end{cases} \tag{4}$$

因此，式(2)中的共现张量$C_{\mathrm{T}}\in R^{M\times N\times D}$可以由$F$与阈值激活张量$A\iota A$卷积得到，记为$C'_{\mathrm{T}}$。这里$A_{\iota A}=A\cdot\iota A$，$\iota A=A>\overline{A}$，$\overline{A}$为张量$A$的均值。为了简化计算，UHFC 将式(3)分成两个不同方向的共现操作，因此，最终的共现张量C_{T}可以由C'_{T}和它的转秩的均值来表示，如式(5)所示：

$$C_{\mathrm{T}}=(C'_{\mathrm{T}}+C'^{\mathrm{T}}_{\mathrm{T}})/2=\{(A_{\iota A}*F)\cdot\iota A+[(A_{\iota A}*F)\cdot\iota A]^{\mathrm{T}}\}/2 \tag{5}$$

2.1.2　特征融合

文献 [7] 中对于提取到的共现特征采用连接的方式将共现向量与图像特征表示符进行融合，但这增加了模型的参数量，使得模型优化困难，尽管采用了 1×1 卷积的方式降低模型的参数量，却使得最终的检索性能下降。Forcen[16] 同样也是采用张量拼接的方式对共现特征进行融合，并提出了双线性池化的方法减少模型的参数量，而传统的特征融合方式并不能充分融合特征的有效信息，造成融合后特征质量不佳。

受 Dai 等[19] 的启发，本文设计了一种适用于共现特征融合的基于空间注意力机制的注意特征融合。在 Dai 等[19] 提出的注意特征融合框架中，加入多尺度通道注意力机制，使得特征在融合过程中减少对干扰通道的关注的同时，保证局部特征不被忽略。

图 3(a) 为 AFF 结构，图 3(b) 为空间注意力机制(SAM)结构。

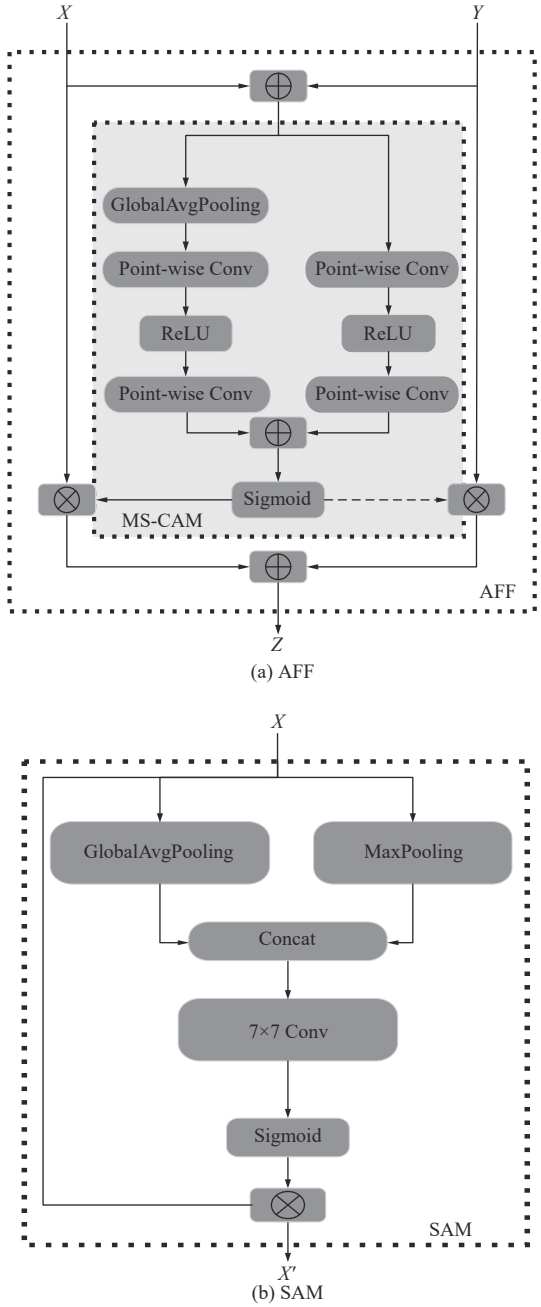

图3　AFF 和 SAM 结构图

Fig. 3 AFF and SAM structure diagram

在本文中，共现特征的计算本身是在不同通道的图特征之间进行的，显然使用多尺度通道注意力机制(MS-CAM)对通道进行选择会使得网络结构冗余，无法有效利用融合的特征信息。对于共现特征与深度特征的融合，抑制背景因素的干扰比通道选择更加有效，能够使得生成的图像特征更加关注图像的主体部分特征。因此，本文将图 3(a)AFF 框架中的 MS-CAM 模块换成了图 3(b) 中的空间注意力机制形成一种新的基于空间注意力的注意特征融合框架(AFF-S)，如图 4 所示。空间注意力特征的输出可以用X'来表示，X表示输入特征，如式(6)所示。

$$X' = X \otimes S(X) = X \otimes \sigma([f^{7\times7}(\mathrm{AvgPool}(X);\mathrm{MaxPool}(X)])) \quad (6)$$

式中：$\boldsymbol{S}(X)$表示经过空间注意力机制的权重向量；⊗表示点乘；$f^{7\times7}$表示卷积核的大小为 7×7；$\sigma()$表示 Sigmoid 函数。

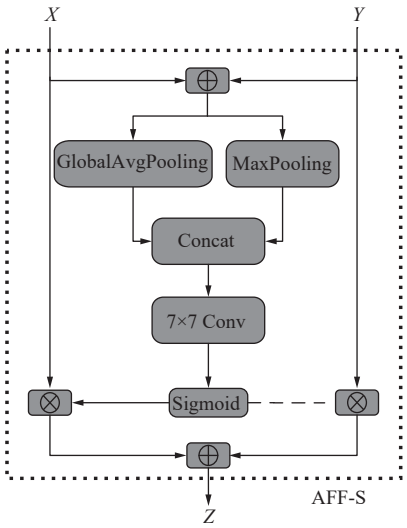

图 4 AFF-S 结构图

Fig. 4 AFF-S structure diagram

因此，原深度特征A与共现特征C_{T}经过 AFF-S 结构融合的特征A_{C}可以表示为

$$A_{\mathrm{C}} = S(A \oplus C_{\mathrm{T}}) \otimes A + (1 - S(A \oplus C_{\mathrm{T}})) \otimes C_{\mathrm{T}} \quad (7)$$

式中：⊕表示张量之间按位相加的操作。

对于输入的两个特征A和C_{T}，AFF-S 首先通过叠加的方式进行合并，并将合并的特征送入 SAM 模块获得特征A的权重向量，由于 Sigmoid 函数的输出在 0 ~ 1 之间，所以这里采用$1-S(A\oplus C_{\mathrm{T}})$的方式获得$C_{\mathrm{T}}$的权重矩阵，实现特征融合过程中权重的自主学习。融合后的张量A_{C}经过池化操作后，展平获得图像在哈希映射前的深度特征$v \in R^{K\times1}$，K为特征的长度。

2.2 无监督哈希算法

哈希映射的过程可以看成两个通道的信息传输问题，Li 等[6]提出当哈希码的分布尽可能均匀时，该哈希码所携带的信息熵能达到最大值。对于M张图片经过 CNN 网络的深度特征$V \in R^{K\times M}$，K代表最后一层特征的长度。$H \in \{1,-1\}^{K\times M}$为$V$经过哈希映射生成的哈希编码。为了实现$V \to H$的有效传输，降低哈希映射过程中的信息损失，本文通过 Li 等[6]提出的双半分布哈希编码构造用于哈希映射过程监督的伪哈希矩阵$B \in \{1,-1\}^{K\times M}$。具体构造方法为：将第$v$个特征位的特征值$v_k \in R^{1\times M}, k \in (1,2,3\cdots K)$,按照特征值大小进行降序排列，并将前半部分特征值赋值为 1，后半部分赋值为−1，来获得该特征位的哈希编码b_k，如式(8)所示：

$$b_k = \begin{cases} 1 & \text{排序后}v_k\text{的前半部分} \\ -1 & \text{其他} \end{cases} \quad (8)$$

对于得到的 B，本文采用$\|\cos(V_1 - V_2) - \cos(B_1 - B_2)\|_2^2$代价函数来最小化 V 与 B 余弦距离关系上的差异，并通过反向传播对 V 的生成进行优化，进而对直接哈希映射 H 进行优化训练。式中 V_1 表示 V 前 $K/2$ 个特征位的特征向量，V_2 表示 V 后 $K/2$ 个特征位的特征向量；B_1、B_2 同理。

文献 [7] 中，为了使得最终的哈希编码保留哈希映射之前的分类信息，通过在哈希映射前后同时添加一层分类层，用 KL 损失函数对输出的分类向量进行拟合，提高最终哈希编码的质量。本文为了避免H与B的过度拟合，提高哈希编码的质量，参照文献 [7] 中的做法，对连续特征V和最终哈希编码H添加一层分类层，分别获得K_1和K_2两个分类矩阵，并采用 KL 散度误差进行优化。因此网络最终的损失函数如式(9)所示：

$$L = L_{\mathrm{mse}} + L_{\mathrm{KL}} = \|\phi - \varphi\|_2^2 + \sum_{i=1}^{M} K_{1_i} \cdot [\lg(K_{1_i}) - \lg(K_{2_i})]/M \quad (9)$$

式中：M为批量训练的图片数；L_{mse}为均方损失；L_{KL}为 KL 损失；ϕ=cos$(V_1 - V_2)$；φ=cos$(B_1 - B_2)$。

3 实验

3.1 数据集

本文在 Flickr25k，Cifar-10，Nus-wide 这 3 个无监督哈希领域常用数据集上进行测试。①Flickr25k 数据集是从社交摄影网站 Flickr 收集的 25 000 张图像，共有 38 个语义标签，属于多标签数据集；本

文随机选择 4 000 张图像作为训练集，1 000 张图片作为测试集，其余图像用作检索的数据库。②Nus-wide-21 数据集包含 195 834 张图像，有 21 个分类；本文在每一个类中随机抽取 500 张图像进行训练，100 张图片用于测试，其余图片用于检索。③Cifar-10 是一个用于识别普适物体的小型数据集，一共包含 10 个类别的 RGB 彩色图片；本文随机选取 5 000 张训练图片和 1 000 张测试图片，并将其余图片作为检索的数据库。

3.2　实验设置

实验在 Windows10 操作系统下，基于 Pytorch 1.12.1 和 Python3.7 在单块 NVIDIA GeForce GTX 1650Ti 的 GPU 环境下进行模型的训练和测试。训练过程中，图片随机裁剪为 256×256 的尺寸，并采用随机翻转和随机裁剪来进行数据增强。优化器采用 Adam 优化器来进行梯度优化，学习率为 0.000 1，权重衰减系数为 5×10^{-4}，Batch size 为 16。在对比实验和消融实验的特征提取部分，共现窗口大小 r 取 4。

3.3　实验结果

3.3.1　对比实验

为验证实验的有效性，本文选取本领域不同阶段的代表模型方法 LSH[1]、SH[20]、ITQ[21]、Deepbit[22]、SGH[23]、Greedy Hash[3]、TBH[24]、HashSIM[8]、DSCH[25] 方法在 Flickr25k、Cifar-10、Nus-wide 数据集上进行对比。

其中，LSH[1]、SH[20]、ITQ[21] 均为传统无监督哈希方法。LSH[1] 即局部敏感哈希，是最初用来做图像索引的哈希算法；SH[20] 为谱哈希，通过对高维数据集进行谱分析，将图像编码过程转换成拉普拉斯特征图的降维问题；ITQ[21] 是一种基于 PCA 的图像哈希方法，通过旋转主成分方向使得各方向的方差尽量保持平衡。

Deepbit[22]、SGH[23]、Greedy Hash[3]、TBH[24]，HashSIM[8]、DSCH[25] 为近年来一些主流的深度哈希方法，其实验数据均来自于相关文献。Deepbit[22] 对二进制码实施了 3 个标准(即最小化量化损失，均匀分布编码和不相关位)学习压缩二进制描述子；SGH[23] 随机生成哈希，通过最短描述符原则来学习哈希函数，使得哈希码可以最大限度地压缩数据集；Greedy Hash 将贪婪思想引入神经网络[3]，解决由于对输出施加离散约束而使得优化变为 NP 难的问题；TBH 在网络中引入了双向瓶颈结构实现对图像的自动编码[24]；HashSIM 通过结构和内在相似性学习(HashSIM)以端到端方式训练哈希编码[8]；DSCH 使用高斯混合模态(GMM) 构建图像的语义组件结构[25]，将图像表示为多个组件的混合。此外，为验证改进的有效性，本文选取 Bif-hash[6] 双半分布哈希模型作为基线模型。实验过程中采用经过预训练的 VGG16 模型作为模型的主干网络，实验结果如表 1 所示。

表 1　对比实验结果

Tab. 1　Comparison with state-of-the art

Method	Flickr25k(mAP@5 000)			Cifar-10(mAP@1 000)			Nus-wide(mAP@5 000)		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bis
LSH	58.3	58.8	59.3	13.2	15.8	16.7	43.2	44.1	44.3
SH	59.1	59.2	60.1	16.0	15.8	15.0	44.6	45.4	49.3
ITQ	61.9	63.2	63.5	19.4	20.9	21.5	52.8	53.2	53.1
Deepbit	59.3	59.3	61.9	22.0	24.1	25.2	45.4	46.2	47.6
SGH	61.6	62.8	62.5	17.9	18.3	18.9	49.4	48.3	48.6
Greedy Hash	62.3	63.1	63.4	28.7	31.7	35.4	51.3	55.8	59.2
TBH	74.3	76.1	77.8	53.2	57.3	57.8	71.7	72.5	73.5
HashSIM	78.8	80.1	80.9	44.7	45.5	45.9	76.6	79.1	80.5
DSCH	81.7	82.7	82.8	**62.4**	**64.4**	**67.0**	77.0	79.2	80.1
Bif-hash (Baseline)	82.7	84.3	85.7	52.4	54.0	55.6	78.7	81.2	81.6
UHFC	**84.7**	**86.9**	**87.8**	54.8	57.8	59.3	**79.2**	**82.1**	**82.8**

由表 1 可以看出，UHFC 的精度在 Flickr25k、Cifar-10、Nus-wide 数据集上依次分别达到了 87.8%，59.3% 和 82.8%，与 Baseline 的精度相比分别提高了 2.1%、3.7%、1.2%，优于现有模型，证明了本文改进的优越性。同时，也证明了在无监督哈希框架中增加一层共现层并通过 AFF-S 的融合方式获得的图像特征可以有效提高最终哈希编码的质量。

值得注意的是，UHFC 在 Cifar-10 数据集上达

到了 59.3% 的精度效果，优于 HashSIM[8] 的，但低于 DSCH[25] 的。原因是 DSCH[25] 通过语义组件的共现构建了粗粒度与细粒度，细粒度与细粒度之间的关系。而哈希在细粒度识别任务上本身会损耗大量精度，所以 UHFC 并没有考虑到图像的细粒度特征。对于 Cifar-10 这种类内相似性较小的数据集 DSCH[25] 更具有优势。且 DSCH 在数据集 Flickr25k 上，当哈希码的长度由 32 bits 变为 64 bits 时，精度只提升了 0.1%，而 UHFC 精度提升了 0.9%，这证明 UHFC 比 DSCH[25] 还有更大的优化空间。此外，表 1 中 16 bits 的 UHFC 和 16 bits 的 Bif-hash[6] 在 Flickr25k 和 Cifar-10 数据集上的性能比 32 bits 的 HashSIM[8] 要好，这得益于 Bif-hash[6] 采用双半分布的伪哈希矩阵，使得哈希码保留更多的特征信息。

3.3.2 消融实验

为了观察引入共现特征以及 KL 损失函数在不同 CNN 框架下，对哈希码性能的影响，本文在 Resnet50，VGG16，MobileNetV3 这 3 个常用的模型上进行了消融实验，采用 Flickr25k 数据集，Batch size 为 16、64 bits，mAP@5000。Baseline 表示使用 CNN 特征和 Bif-Hash 方法生成哈希编码，KL+表示在生成哈希编码的过程中引入 KL 散度损失，Cooc+表示采用融合共现特征的深度特征，Cooc*表示采用改进的共现特征，UHFC 表示同时采用特征优化和损失优化，实验结果如表 2 所示。

表 2 消融实验结果

Tab. 2 Ablation experiment result

Method	ResNet50	VGG16	MobileNetV3
Baseline	85.3	85.7	83.6
KL+	85.9	86.3	84.1
Cooc+	86.3	86.8	84.3
Cooc*	86.9	87.3	84.7
UHFC	**87.4**	**87.8**	**85.1**

从表 2 可以看到，与 Baseline 相比本文方法在 ResNet50，VGG16，MobileNetV3 网络上，分别有 1.9%、2.1%、1.5% 的提升，说明本文方法能够适用于不同的卷积网络框架。另一方面，表 2 中 Cooc+要比 KL+的精度高，证明了对哈希映射前的特征进行优化比对损失函数的优化更有利于提高哈希码的质量。同时表中 Cooc*的检索精度高于 Cooc+的证明了采用激活值的均值来表示特征的共现程度能够提高深度特征的质量。最后在反向传播中引入 KL 损失函数，在 Resnet50 上相比于 Baseline 提高了 0.6% 的精度证明了在哈希映射前后添加分类层可以提高哈希码的检索精度，验证了本文方法的有效性。

3.3.3 融合方法对比

在第 2.1.2 节中，UHFC 设计了适用于共现特征融合的注意特征融合框架 AFF-S，本节为了验证 AFF-S 的有效性，采用 VGG16 作为 CNN 的基本框架，在 Flickr25k 数据集上对不同特征融合方法的 mAP@5000 进行了比较。实验结果如表 3 所示，表中 Linear fusion 为融合方法对比实验的基线方法，表示采用线性融合的方式将深度特征及共现特征按 1：1 的比例进行融合；AFF 为 Dai 等[19] 提出的基于多尺度通道注意力机制的注意力融合方法，此外 Dai 等[19] 还设计了一个嵌套的注意特征融合框架即 iAFF；Linear+MACAM 表示在线性融合后的特征后增加 MS-CAM 模块；AFF-S 表示本文提出的基于空间注意力的特征融合机制。

表 3 融合方法对比实验结果

Tab. 3 Experimental results of fusion method compares

融合方式	16 bits	32 bits	64 bits
Linear fusion	84.1	85.5	86.3
AFF	75.1	77.8	86.0
iAFF	83.8	85.9	86.7
Linear+MSCAM	84.2	85.7	86.7
AFF-S	**84.7**	**86.9**	**87.8**

从表 3 可以看出，AFF-S 在不同长度的哈希码上表现都优于其他特征融合方法，证实了 AFF-S 框架对于共现特征融合的有效性。此外，表 3 中无论是 AFF 还是 iAFF 相对于 Linear fusion 性能都有不同程度的下降，而 Linear+MSCAM 的性能相比于 Linear fusion 有微小的提升。这表示 MS-CAM 模块本身并不适用于共现特征的融合，但能够对融合后的特征进行优化。原因是因为共现特征本身是对不同通道内特征之间的操作，MS-CAM 模块同样也是对不同通道的特征分配权重的，采用 MS-CAM 模块对特征进行融合会造成大量通道信息的丢失，降低模型性能。AFF-S 的实验结果也验证了将 AFF 框架中的 MS-CAM 模块换成空间注意力机制的正确性。

3.3.4 参数敏感性分析

在第 2.1.1 节中，共现定义为不同通道的两个

激活值均大于 s，且同在长度为 $2r-1$ 窗口下。为了减少训练过程中的参数量，在计算共现特征时，本文将 s 设置为最后一层卷积输出的激活张量的均值，r 设置为固定值。在本节，为了观察 r 对模型性能的影响，进行了参数敏感性分析。由图 5 可以看到，共现参数 $r=4$ 时效果最好。原因在于大一点的共现窗口能有效捕捉到部分之间的联系，过小的共现窗口会捕捉到一些噪声干扰，不利于图像检索，造成检索性能降低。而当 $r=5$ 时，检索的性能降低，由此可以得出过大的共现窗口会造成信息丢失，也会导致精度的下降。

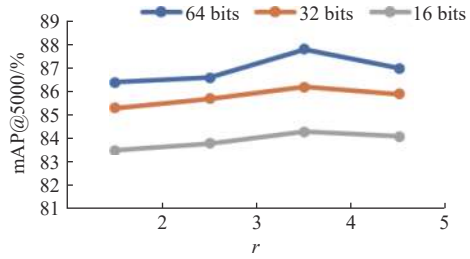

图 5　共现窗口参数 r 实验

Fig. 5 Co-occurrence window parameter r experiment

4　结束语

在基于深度特征的无监督哈希检索研究中，图像特征的质量对检索的精度具有决定性作用。为提高最终哈希检索结果的精度，本文在 CNN 网络中增加了一层共现层，用来提取深度特征部分之间的依赖关系，并设计了一种用于共现特征融合的注意特征融合框架 AFF-S，通过特征融合生成最终的深度特征。实验证明在最后一层卷积特征中融合共现特征可以有效提高深度特征的质量，提高哈希检索的精度。此外，为了提高哈希码所携带的图像信息，本文在哈希映射前后分别增加一层分类层，通过分类层输出之间的 KL 损失对哈希映射进行优化。本文方法在 Flickr25k 和 Nus-wide 数据集上获得 87.8% 和 82.3% 的精度，优于现有无监督哈希方法，验证了本文方法的有效性。最后，本文为了减少网络的参数，只在最后一层卷积层后增加了共现层，且整体模型在类内差异较少的数据集上的表现略低于当前先进模型。未来工作中会尝试将共现层应用在卷积网络其他层的输出并结合图像的细粒度特征以达到更高的性能。

参考文献:

[1] DATAR M, IMMORLICA N, INDYK P, et al. Locality-sensitive hashing scheme based on p-stable distributions[C]//Proceedings of the Twentieth Annual Symposium on Computational Geometry. New York: ACM, 2004: 253-262. DOI: 10.1145/997817.997857.

[2] LI W J, WANG S, KANG W C. Feature learning based deep supervised hashing with pairwise labels[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: IJCAI/AAAI Press, 2016: 1711-1717.

[3] SU S P, ZHANG C, HAN K, et al. Greedy hash: towards fast optimization for accurate hash coding in CNN[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc. , 2018: 806-815.

[4] CAO Z J, LONG M S, WANG J M, et al. HashNet: deep learning to hash by continuation[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 5608-5617. DOI: 10.1109/ICCV.2017.598.

[5] MA Y, LI Q, SHI X S, et al. Unsupervised deep pairwise hashing[J]. Electronics, 2022, 11(5): 744. DOI: 10.3390/electronics11050744.

[6] LI Y, Q VAN GEMERT J. Deep unsupervised image hashing by maximizing bit entropy[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 2002-2010. DOI: 10.1609/aaai.v35i3.16296.

[7] GONG Q K, WANG L D, LAI H J, et al. ViT2Hash: unsupervised information-preserving hashing[J]. arXiv preprint arXiv, 2022: 2201.05541.

[8] LUO X, MA Z Y, CHENG W, et al. Improve deep unsupervised hashing via structural and intrinsic similarity learning[J]. IEEE Signal Processing Letters, 2022, 29: 602-606. DOI: 10.1109/LSP.2022.3148674.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90. DOI: 10.1145/3065386.

[10] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[11] WU Z B, YU J Q. A multi-level descriptor using ultra-deep feature for image retrieval[J]. Multimedia Tools and Applications, 2019, 78(18): 25655-25672. DOI: 10.1007/s11042-019-07771-2.

[12] ZHANG Z L, ZHANG X Y, PENG C, et al. ExFuse: enhancing feature fusion for semantic segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 269-284. DOI: 10.1007/978-3-030-01249-6_17.

[13] DAMANEH M M, MOHANNA F, JAFARI P. Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter[J]. Expert Systems with Applications, 2023, 211: 118559. DOI: 10.1016/j.eswa.2022.118559.

[14] YANG Y, NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2010: 270-279. DOI: 10.1145/1869790.1869829.

[15] ZHU W T, LAN C L, XING J L, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016. DOI: 10.1609/aaai.v30i1.10451.

[16] FORCEN J I, PAGOLA M, BARRENECHEA E, et al. *Co-occurrence* of deep convolutional features for image search[J]. Image and Vision Computing, 2020, 97: 103909. DOI: 10.1016/j.imavis.2020.103909.

[17] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1-9. DOI: 10.1109/CVPR.2015.7298594.

[18] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2117-2125. DOI: 10.1109/CVPR.2017.106.

[19] DAI Y M, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2021: 3560-3569. DOI: 10.1109/WACV48630.2021.00360.

[20] WEISS Y, TORRALBA A, FERGUS R. Spectral hashing[C]//Proceedings of the 21st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc. , 2008: 1753-1760.

[21] GONG Y C, LAZEBNIK S, GORDO A, et al. Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2916-2929. DOI: 10.1109/TPAMI.2012.193.

[22] LIN K, LU J W, CHEN C S, et al. Learning compact binary descriptors with unsupervised deep neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1183-1192. DOI: 10.1109/CVPR.2016.133.

[23] DAI B, GUO R Q, KUMAR S, et al. Stochastic generative hashing[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017: 913-922.

[24] SHEN Y M, QIN J, CHEN J X, et al. Auto-encoding twin-bottleneck hashing[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2818-2827. DOI: 10.1109/CVPR42600.2020.00289.

[25] LIN Q H, CHEN X J, ZHANG Q, et al. Deep unsupervised hashing with latent semantic components[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 7488-7496. DOI: 10.1609/aaai.v36i7.20713.

作者简介：

张青露 硕士, zqlkdbb@foxmail.com

陈壹华(通信作者) 副教授, chenyihua@m.scnu.edu.cn